The Rate of Convergence of a Matrix Power Series

N. J. Young
Department of Mathematics
The University
Glasgow, Scotland

Submitted by Stephen Barnett

ABSTRACT

To estimate the truncation error of a matrix power series we need information about the magnitude of high powers of a matrix. Inequalities bearing on this question are surveyed, and their use is exemplified by calculations of bounds for the truncation error of the geometric series in matrices.

INTRODUCTION

Iterative methods for solving problems involving matrices sometimes correspond to the summation of a matrix power series. An obvious example is the inversion of a matrix using a geometric series:

$$(I-A)^{-1}=I+A+A^2+\cdots. \tag{1}$$

A less trivial example, illustrating the fact that the coefficients may be matrices, is the solution of the Lyapunov matrix equation

$$X-A^*XA=B$$

by the iterative summation of the series [10]

$$X=B+A^*BA+A^{*2}BA^2+\cdots. \tag{2}$$

Here A^* is the conjugate transpose of A.

Satisfactory sufficient conditions for series such as (1) and (2) to converge can often be deduced quite simply from the spectral-radius formula. This tells us that, if $\|\cdot\|$ is any matrix norm on the algebra of $n\times n$ complex

52 Vanderbilt Ave., New York, NY 10017 0024-3795/81/010261+18$02.50

matrices, then

$$\lim_{m\to\infty} \| A^m \|^{1/m} = |A|_\sigma,$$

where $|A|_\sigma$ is the spectral radius of A, that is, $\max\{|\lambda|: \lambda \text{ is an eigenvalue of } A\}$. This means that $\| A^m \|$ behaves asymptotically like $|A|_\sigma^m$ as $m\to\infty$, and it follows that (1) and (2) converge absolutely when $|A|_\sigma < 1$.

Asymptotic estimates of rates of convergence can also be inferred from the spectral radius formula. If we write

$$Y_k = I + A + \cdots + A^{k-1},$$

then the error incurred in stopping the summation (1) at the kth term is

$$\begin{aligned}(I-A)^{-1} - Y_k &= A^k + A^{k+1} + \cdots \\ &= A^k (I-A)^{-1}. \end{aligned} \tag{3}$$

Hence

$$\begin{aligned}\| (I-A)^{-1} - Y_k \|^{1/k} &\leqslant \| A^k \|^{1/k} \| (I-A)^{-1} \|^{1/k} \\ &\to |A|_\sigma \qquad \text{as} \quad k\to\infty.\end{aligned}$$

Likewise, if $X_k = \sum_{m=0}^{k-1} A^{*m} B A^m$,

$$\begin{aligned}\| X - X_k \|^{1/k} &= \| A^{*k} X A^k \|^{1/k} \leqslant \left(\| X \| \, \| A^{*k} \| \, \| A^k \| \right)^{1/k} \\ &\to |A|_\sigma^2 \qquad \text{as} \quad k\to\infty. \end{aligned} \tag{4}$$

However, this type of information is very weak for practical applications. In summing (2) we might perform a dozen iterations, which corresponds to evaluating X_k with $k = 2^{12}$ [10], and knowledge of the limiting value of $\| X - X_k \|^{1/k}$ is only a very slight indication of what kind of accuracy we can expect in advance. In particular it tells us nothing of the role played by the size of the matrices. This paper is devoted to ways of obtaining better estimates.

The formulae (3) and (4) show that a useful step would be to find an upper bound for $\| A^m \|$. Of course, for a matrix norm $\| A^m \| \leqslant \| A \|^m$, so that if $\| A \| < 1$ we can immediately estimate the rates of convergence in (1) and

(2). But a satisfactory estimate for $\|A^m\|$ must have the right asymptotic behavior, and must therefore involve the spectral radius. This leads to the problem of finding the supremum of $\|A^m\|$ for given values of $\|A\|$ and $|A|_\sigma$.

In the particular case that $\|\cdot\|$ is the operator norm on n-dimensional Hilbert space and $m=n$, quite a lot is known. This case was originally raised by V. Pták in [6]: it was motivated by considerations arising from his notion of critical exponent. A summary of the results and methods relating to this problem can be found in [9]. In the present context we are more interested in the case that m is considerably larger than n, but nevertheless some of the ideas described in [9] will prove useful.

Since all norms on the algebra of $n\times n$ complex matrices are equivalent, one should get qualitatively similar results whichever norm one uses. However, we are interested in actual (nonasymptotic) bounds, and so it does make a difference which we choose. For clarity we shall use $\|\cdot\|$ to denote any operator norm (that is, any norm which is submultiplicative and satisfies $\|I\|=1$) and reserve $|\cdot|$ for the operator norm on n-dimensional Hilbert space: thus, if $x=[x_1\cdots x_n]^T$ we write $|x|=\{\sum_{i=1}^n |x_i|^2\}^{1/2}$ and $|A|=\sup\{|Ax|:|x|\leqslant 1\}$. It is well known that $|A|=|A^*A|_\sigma^{1/2}$; $|\cdot|$ is sometimes called the spectral norm. And $|\cdot|_p$ denotes the Schatten–von Neumann norm:

$$|A|_p=\left[\operatorname{trace}(A^*A)^{p/2}\right]^{1/p}, \qquad 0<p<\infty$$

(see [4]). In particular, $|\cdot|_1$ is the trace norm and $|\cdot|_2$ is the Euclidean or Hilbert-Schmidt norm.

The three main methods of estimating norms of functions of matrices, in order of decreasing simplicity, can be described as (1) bare hands and induction; (2) the quotient norm method; (3) the geometry of Hilbert space. They are described in numerous papers; the purpose of this article is to bring them together and illustrate their relevance to the estimation problem outlined above. In Secs. 1 to 3 the norm inequalities are presented, with some improvements on published results (in Theorem 1 and the example in Sec. 3). In Sec. 4 these inequalities are applied to the accelerated summation of the geometric series (1). It is shown, for example, that even for a 50×50 matrix of norm 1 having an eigenvalue of magnitude 0.9, the truncation error is negligible after 12 iterations.

1. BARE HANDS AND INDUCTION

Two methods come under this heading. The first is the simplest and neatest of any. The idea is due to C. Apostol.

THEOREM 1. *Let $T_1, \dots, T_m$ be upper triangular $n \times n$ matrices, $m \geqslant n$, let $|T_i| \leqslant 1$, $1 \leqslant i \leqslant m$, and let the jth entry on the main diagonal of T_i have modulus no greater than r_j, $1 \leqslant i \leqslant m$, $1 \leqslant j \leqslant n$. Then*

$$|T_1 T_2 \cdots T_m|_1 \leqslant h_{m-n+1}(r_1, \dots, r_n), \tag{5}$$

where $h_k(r_1, \dots, r_n)$ denotes the sum of all monomials in $r_1, \dots, r_n$ of degree k.

Dr. Apostol communicated the case $m = n$ orally, and his proof is given in [9, Sec. 4].

Proof. Certainly (5) holds if $n = 1$. Consider a pair m, n with $m \geqslant n$, and suppose that (5) is true for any pair m', n' such that $m' \geqslant n'$ and $m' + n' < m + n$. Let

$$T_i = \begin{bmatrix} R_i & * \\ 0 & \lambda_i \end{bmatrix}.$$

Then R_i is of type $(n-1) \times (n-1)$, $|R_i| \leqslant 1$, and $|\lambda_i| \leqslant r_n$. We have

$$\begin{aligned} T_1 T_2 \cdots T_m &= (T_1 \cdots T_{m-1}) T_m \\ &= \begin{bmatrix} R_1 \cdots R_{m-1} & * \\ 0 & \lambda_1 \cdots \lambda_{m-1} \end{bmatrix} \begin{bmatrix} R_m & * \\ 0 & \lambda_m \end{bmatrix} \\ &= \begin{bmatrix} R_1 \cdots R_{m-1} & 0 \\ 0 & 0 \end{bmatrix} T_m + T_1 \cdots T_{m-1} \begin{bmatrix} 0 & 0 \\ 0 & \lambda_m \end{bmatrix}. \end{aligned}$$

Since $|ABC|_1 \leqslant |A|\,|B|_1\,|C|$,

$$|T_1 \cdots T_m|_1 \leqslant |R_1 \cdots R_{m-1}|_1 |T_m| + |T_1 \cdots T_{m-1}|\,|\lambda_m|.$$

Now if $m - 1 \geqslant n$ the inductive hypothesis yields

$$|T_1 \cdots T_{m-1}| \leqslant h_{m-n}(r_1, \dots, r_n),$$

since $|A| \leqslant |A|_1$ for all A. If we define $h_0(r_1, \dots, r_n)$ to be identically 1, this remains true for the case $m = n$, and on applying the inductive hypothesis a second time we have

$$\begin{aligned} |T_1 \cdots T_m|_1 &\leqslant h_{m-n+1}(r_1, \dots, r_{n-1}) + r_n h_{m-n}(r_1, \dots, r_n) \\ &= h_{m-n+1}(r_1, \dots, r_n). \end{aligned}$$

Now consider an arbitrary $n \times n$ matrix A. Pick a unitary matrix U such that U^*AU is upper triangular, and apply Theorem 1 with $T_i = (U^*AU)/|A|$, $r_i = |A|_\sigma/|A|$, $1 \leqslant i \leqslant m$. Since $|\cdot|_1$ is invariant with respect to unitary similarity, we obtain

$$\frac{|A^m|_1}{|A|^m} \leqslant h_{m-n+1}(r_1, r_2, \ldots, r_n).$$

The number of monomials of degree k in n variables is $\binom{k+n-1}{n-1}$, and so, if $r_1 = r_2 = \cdots = r_n = r$,

$$h_{m-n+1}(r_1, \ldots, r_n) = \binom{m}{n-1} r^{m-n+1}.$$

We therefore have the following neat conclusion.

COROLLARY. *For any $n \times n$ matrix A,*

$$|A^m| \leqslant |A^m|_1 \leqslant \binom{m}{n-1} |A|^{n-1} |A|_\sigma^{m-n+1}. \tag{6}$$

The second induction method, which is essentially due to J. D. Stafney [11], applies to more general functions of A. In fact Stafney's calculations contain an error and the inequality he gives (Theorem 2.1 of [11]) is incorrect, but his reasoning can be modified to give the following

THEOREM 2. *If A is an $n \times n$ matrix, f is a function analytic in a convex open set containing the eigenvalues of A, and $0 < p \leqslant 2$, then*

$$|f(A)|_p \leqslant \left\{ \sum_{j=0}^{n-1} \binom{n}{j+1} \frac{|A|^{pj}}{(j!)^p} \|f^{(j)}\|_A^p \right\}^{1/p}, \tag{7}$$

where $\|f\|_A$ denotes the maximum of $|f|$ on the convex hull of the eigenvalues of A.

A detailed proof of this is given in [13]; here is a sketch. As above, we can reduce to the case of upper triangular A using a unitary similarity, and so can write

$$A = \begin{bmatrix} B & w \\ 0 & \alpha \end{bmatrix},$$

where B is one dimension smaller. It follows from the definition of $f(A)$ in terms of the Cauchy integral formula that

$$f(A)=\begin{bmatrix} f(B) & g(B)w \\ 0 & f(\alpha) \end{bmatrix},$$

where $g(z)=(f(z)-f(\alpha))/(z-\alpha)$. The inequality (7) can be established by induction in view of the facts that, if $0<p\leqslant 2$,

$$|f(A)|_p^p \leqslant |f(B)|_p^p + |g(B)w|_p^p + |f(\alpha)|^p$$

and

$$\|g^{(j)}\|_B \leqslant \frac{1}{j+1}\|f(j+1)\|_A.$$

Theorem 2 can of course be used to estimate $|A^m|$, but the result is strictly worse than the Corollary to Theorem 1; the right-hand side of (6) is in fact precisely the term of lowest degree in $|A|_\sigma$ when one substitutes $f(z)=z^m$ in (7).

We note that (6) does give the right order of magnitude for $|A^m|$ as $|A|_\sigma \to 0$. We shall see later that there exists an $n\times n$ matrix A such that $|A|=1$, $|A|_\sigma=r<1$, and

$$|A^m|=\binom{m}{n-1}r^{m-n+1}+O(r^{m-n+2}) \qquad \text{as} \quad r\to 0.$$

2. THE QUOTIENT NORM METHOD

Suppose that $\|\cdot\|_0$ is a norm on the algebra $\mathbf{C}[z]$ of complex polynomials, and that $\|\cdot\|$ is an operator norm such that $\|f(A)\|\leqslant\|f\|_0$ for all $f\in\mathbf{C}[z]$ and all A satisfying $\|A\|\leqslant 1$. We may then be able to estimate $\|f(A)\|$ in terms of $|A|_\sigma$ in the following way.

Let A be of type $n\times n$, $\|A\|\leqslant 1$, $|A|_\sigma\leqslant r$. By the Cayley-Hamilton theorem there is a polynomial

$$p(z)=(z-\alpha_1)\cdots(z-\alpha_n) \tag{8}$$

with $|\alpha_i|\leqslant r$, $1\leqslant i\leqslant n$, such that $p(A)=0$. For any polynomial g, $f(A)=(f+gp)(A)$ and hence

$$\|f(A)\|\leqslant\inf\{\|f+gp\|_0 : g\in\mathbf{C}[z]\}.$$

The right hand side is, by definition, the quotient norm induced by $\|\cdot\|_0$ on the factor space $\mathbf{C}[z]/p\mathbf{C}[z]$; we denote it by $\|f+(p)\|_0$. If we can show that $\|f+(p)\|_0$ is no greater than some constant $K(r)$ for all choices of p such that $|\alpha_i| \leqslant r$, it will follow that $\|f(A)\| \leqslant K(r)$.

It is a surprising fact that this simple idea yields nontrivial information for an arbitrary operator norm $\|\cdot\|$, for if we define $\|\cdot\|_0$ on $\mathbf{C}[z]$ by

$$\|c_0+c_1z+\cdots+c_kz^k\|_0=|c_0|+|c_1|+\cdots+|c_k|, \tag{9}$$

then we clearly have $\|f(A)\| \leqslant \|f\|_0$ whenever $\|A\| \leqslant 1$. This observation was first used in the context of estimating norms of matrix powers by Z. Dostál [1], and the idea was further elaborated by Dostál [2], Pták [8], and Young [12, 14]. The most general statement is the following [14].

THEOREM 3. *Let A be an $n\times n$ matrix such that $|A|_\sigma \leqslant r$. Let $f(z)=\sum_0^\infty a_m z^m$ be analytic on $\{z:|z| \leqslant r\}$, and suppose that $a_m \leqslant 0$ whenever $m<n$ and $n-m$ is even, $a_m \geqslant 0$ for all other m. Then, for any operator norm $\|\cdot\|$,*

$$\|f(A)\| \leqslant \sum_{j=0}^{n-1} \frac{(-1)^{n-j-1}}{j!}(r+\|A\|)^j f^{(j)}(r). \tag{10}$$

The assertion actually remains good for any algebraic element A of degree n in an arbitrary Banach algebra with identity of unit norm. It is sharp in the sense that, for any given n, r, and f, one can find A and $\|\cdot\|$ such that $|A|_\sigma \leqslant r$ and (10) holds with equality.

We can gain some appreciation of the meaning of (10) by putting $f(z)=z^m$ $(m \geqslant n)$. The result makes for interesting comparison with Theorems 1 and 2.

COROLLARY. *If A is of type $n\times n$ and $|A|_\sigma \leqslant r$, then*

$$\|A^m\| \leqslant \sum_{j=0}^{n-1} (-1)^{n-j-1}\binom{m}{j}(r+\|A\|)^j r^{m-j}.$$

The term of smallest degree in r here is given by $j=n-1$, and we find that if $\|A\| \leqslant 1$,

$$\|A^m\| \leqslant \binom{m}{n-1} r^{m-n+1}+O(r^{m-n+2}),$$

which is to say that almost the same holds for an arbitrary operator norm as does for the special norms $|\cdot|$ and $|\cdot|_p$.

The simplest way to prove Theorem 3 is to express the remainder on dividing f by the polynomial $(z-\alpha_1)\cdots(z-\alpha_n)$ by means of a contour integral. This remainder is a polynomial of degree $n-1$ in z whose coefficients are power series in $\alpha_1,\ldots,\alpha_n$, and a little manipulation shows that the coefficients in each of these power series all have the same sign. This enables us to conclude that the sum of the absolute values of the coefficients of the z^k as α_i varies subject to $|\alpha_i|\leqslant r$ is greatest when each $\alpha_i=r$. The inequality (10) then follows simply.

The use of the norm $\|\cdot\|_0$ of (9) in the quotient norm method looks a bit crude, and we could expect to do better using different norms. Indeed, there is a result known as von Neumann's inequality which looks particularly promising in this context. It asserts that, if A is an $n\times n$ matrix and $|A|\leqslant 1$, then for any polynomial f,

$$|f(A)|\leqslant \|f\|_{H^\infty},$$

where

$$\|f\|_{H^\infty}=\sup_{|z|<1}|f(z)|$$

(see [5]). It follows that if $|A|\leqslant 1$ and $p(A)=0$,

$$|f(A)|\leqslant \|f+(p)\|_{H^\infty/(p)} \tag{11}$$

and results of Pták, Sz.-Nagy, and Sarason (see [9]) show that (11) does hold with equality for some A satisfying $|A|=1$, $p(A)=0$, so that the inequality is sharp. This reduces the problem of estimating $|f(A)|$ to a purely function-theoretic question. This looks like an advance, but to date it has been disappointingly unfruitful. The one application of (11) to our problem [7] gives a weaker result than the simple calculation in Theorem 1. At the moment it looks as though there is a better prospect of progress in the opposite direction: certain classical interpolation problems admit a reformulation in terms of matrix extremal problems, which furnishes new insights into them. Such questions are discussed in [9, Sec. 7] and [3].

3. GEOMETRY OF HILBERT SPACE

The Corollary to Theorem 1 shows us that if $|A|\leqslant 1$ and $|A|_\sigma\leqslant r$, where A is of type $n\times n$, then

$$|A^m|\leqslant\binom{m}{n-1}r^{m-n+1}.$$

In the applications we envisage we wish to know that $|A^m|$ is small: this will be so if, for fixed $r<1$, m is sufficiently large or, for fixed m, r is sufficiently small. For quite a lot of values of m and r, however, the bound we obtain is not even less than one, so that the estimate gives us no information. This objection also applies to the estimates in Theorems 2 and 3. Other, and more delicate, methods seem to be needed to ascertain how large $|A^m|$ can be when $|A|_\sigma$ is close to $|A|$ and m is of moderate size. Results of this nature were obtained by Young [15, 16], building on work of Pták [6], by using rather intricate matrix manipulations. Subsequently a much simpler geometric approach was discovered by Pták and Young [9]. In this the problem is reduced to the estimation of the norm of a certain finite rank operator in the Hilbert space H^2 of analytic functions in the open unit disc. In this instance using functions rather than matrices turns out to be an improvement, and the desired norm is estimated by operating with bases of certain finite-dimensional subspaces of H^2.

THEOREM 4. *If A is of type $n\times n$, $|A|\leqslant 1$, and $|A|_\sigma\leqslant r$, then*

$$|A^n|^2\leqslant 1-\left\{\sum_{k=0}^{n-1}\binom{2k}{k}\right\}^{-2}(1-r^2)^{2n-1}.$$

See [9, Sec. 5]. Since $|A^m|\leqslant|A^n|$ when $m\geqslant n$, this shows that the largest possible value of $|A^m|$, subject to $|A|\leqslant 1$, $|A|_\sigma\leqslant r$, is $1-O\big((1-r)^{2n-1}\big)$. One might expect to do better if m is much larger than n; that is, one might expect the maximum of $|A^m|$ to be $1-O\big((1-r)^k\big)$ where $k<2n-1$. Here is an example to confound such an expectation.

Fix r, $0<r<1$. Let H_n be the n-dimensional space consisting of all rational functions of the form $p(z)/(1-rz)^n$ where p is a polynomial of degree less than n. H_n becomes a Hilbert space if we define an inner product by

$$(f,g)=\frac{1}{2\pi i}\int f(z)\overline{g(z)}\,\frac{dz}{z},$$

the integral being taken anticlockwise around the unit circle.

Define $S\colon H_n\to H_n$ by

$$Sf(z)=\frac{1}{z}\big[f(z)-f(0)\big].$$

Then $|S|=1$: this follows from the facts that, if

$$f(z)=a_0+a_1z+\cdots,$$

then

$$|f|^2=(f,f)=|a_0|^2+|a_1|^2+\cdots$$

and

$$Sf(z)=a_1+a_2z+\cdots$$

If p is of degree $\leqslant n-1$,

$$(S-rI)\frac{p(z)}{(1-rz)^{n-1}}=\frac{p(z)-(1-rz)^{n-1}p(0)}{z(1-rz)^{n-1}}\in H_{n-1}.$$

Thus $(S-rI)H_n\subseteq H_{n-1}$, and so $(S-rI)^nH_n=\{0\}$. S consequently has the unique eigenvalue r, and $|S|_\sigma=r$,

We wish to show that if $m\geqslant n$, $|S^m|\geqslant 1-O\big((1-r)^{2n-1}\big)$. Let $g(z)=z^{n-1}/(1-rz)^n$. Then

$$\begin{aligned}
|g|^2&=\frac{1}{2\pi i}\int\frac{z^{n-1}}{(1-rz)^n}\frac{\bar z^{n-1}}{(1-r\bar z)^n}\frac{dz}{z}\\
&=\frac{1}{2\pi i}\int\frac{z^{n-1}dz}{(1-rz)^n(z-r)^n}\\
&=\frac{1}{(n-1)!}\frac{d^{n-1}}{dz^{n-1}}\frac{z^{n-1}}{(1-rz)^n}\bigg|_{z=r}\\
&=\frac{1}{(n-1)!}\sum_{j=0}^{n-1}\binom{n-1}{j}\frac{n(n+1)\cdots(n+j-1)r^j}{(1-rz)^{n+j}}\\
&\quad\times(n-1)(n-2)\cdots(j+1)z^j\bigg|_{z=r}\\
&=\sum_{j=0}^{n-1}\binom{n-1}{j}\binom{n+j-1}{n-1}\frac{r^{2j}}{(1-r^2)^{n+j}}\\
&=\binom{2n-2}{n-1}\frac{r^{2n-2}}{(1-r^2)^{2n-1}}+O\big((1-r^2)^{-2n+2}\big)\qquad\text{as } r\to 1\\
&=\binom{2n-2}{n-1}\frac{2^{-2n+1}}{(1-r)^{2n-1}}+O\big((1-r)^{-2n+2}\big)\qquad\text{as } r\to 1.
\end{aligned}$$

If we write

$$g(z)=c_0+c_1z+\cdots,$$

then

$$S^mg(z)=c_m+c_{m+1}z+\cdots$$

and hence

$$\begin{aligned}|S^mg|^2&=|c_m|^2+|c_{m+1}|^2+\cdots\\&=|g|^2-\left(|c_0|^2+|c_1|^2+\cdots+|c_{m-1}|^2\right).\end{aligned}$$

In fact we have

$$g(z)=z^{n-1}+\binom{n}{1}rz^n+\binom{n+1}{2}r^2z^{n+1}+\cdots,$$

so that, if $m\geqslant n$,

$$\begin{aligned}|c_0|^2+\cdots+|c_{m-1}|^2&=1+\binom{n}{1}^2r^2+\cdots+\binom{m-1}{m-n}^2r^{2(m-n)}\\&\leqslant(m-n+1)\binom{m-1}{n-1}^2.\end{aligned}$$

Hence

$$|S^mg|^2\geqslant|g|^2-(m-n+1)\binom{m-1}{n-1}^2,$$

and so

$$\begin{aligned}|S^m|^2\geqslant\frac{|S^mg|^2}{|g|^2}&\geqslant1-\frac{m-n+1}{|g|^2}\binom{m-1}{n-1}^2\\&=1-(m-n+1)\binom{m-1}{n-1}^2\binom{2n-2}{n-1}^{-1}2^{2n-1}(1-r)^{2n-1}\\&\qquad+O\left((1-r)^{2n}\right)\qquad\text{as}\quad r\to1.\end{aligned}$$

It follows that

$$1-|S^m|=O\big((1-r)^{2n-1}\big) \qquad \text{as} \quad r\to 1,$$

confirming the above claim.

This example, together with Theorem 4, gives us quite a good idea of the behaviour of the quantity

$$C=\sup\{|A^m|: A \text{ is } n\times n,\ |A|\leqslant 1,\ |A|_\sigma \leqslant r\}.$$

If we hold m and n fixed and plot C against r (where $m\geqslant n$, r goes from 0 to 1) we get a curve which has $m-n$ zero derivatives at $r=0$, then ascends rapidly so that it has $2n-2$ zero derivatives at $r=1$.

Table 1, reprinted from [15], illustrates this description. It shows the value of C corresponding to the indicated values of n and r in the case $m=n$. It would seem that m must be substantially bigger than n if it is to be possible to deduce that $|A^m|$ is small from such information as, say $|A|\leqslant 1$, $|A|_\sigma \leqslant \frac{1}{2}$.

The operator S also enables us to establish the sharpness of the estimate (6) as $|A|_\sigma \to 0$. It is shown in [9] that the matrix of S with respect to a certain orthonormal basis of H_n is $(N+rI)(I+rN)^{-1}$, where N is the $n\times n$ shift matrix

$$N=\begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

TABLE 1

r \ n	2	3	4	5	10	15	20
.1	.19850377	.29407320	.38538375	.47127467	.79467159	.94248736	.98761933
.2	.38812241	.55432685	.69189744	.79825775	.98828198	.99965068	.99999153
.3	.56045260	.75542086	.88003587	.94775371	.99970488	.99999903	1.0000000
.4	.70815051	.88747509	.96511898	.99081806	.99999537	1.0000000	1.0000000
.5	.82569391	.95844098	.99258334	.99887345	.99999996	1.0000000	1.0000000
.6	.91036109	.98834583	.99888906	.99990752	1.0000000	1.0000000	1.0000000
.7	.96325820	.99775152	.99989670	.99999577	1.0000000	1.0000000	1.0000000
.8	.98981479	.99976605	.99999581	.99999993	1.0000000	1.0000000	1.0000000
.9	.99885325	.99999429	.99999998	1.0000000	1.0000000	1.0000000	1.0000000

It follows that the matrix of S^m is

$$(N+rI)^m(I+rN)^{-m}=\left\{\sum_{j=0}^{m}\binom{m}{j}N^j r^{m-j}\right\}\left\{\sum_{k=0}^{\infty}\binom{m+k-1}{k}r^k N^k\right\}.$$

Since $N^n=0$, $N^{n-1}\neq 0$, the lowest power of r here is given by $j=n-1$, $k=0$. Thus

$$(N+rI)^m(I+rN)^{-m}=\binom{m}{n-1}r^{m-n+1}N^{n-1}+O(r^{m-n+2}).$$

Hence

$$|S^m|\geqslant|S^m|_1=\binom{m}{n-1}r^{m-n+1}+O(r^{m-n+2}).$$

4. THE TRUNCATION ERROR OF THE GEOMETRIC SERIES

Which of the foregoing results will be most effective for our original purpose, namely, gaining some idea of how many terms of a power series we need to sum? Let us try them out on the geometric series (1). As before let Y_k be the sum of the first k terms. Denote the truncation error $(I-A)^{-1}-Y_k$ by E_k, and let $|A|_\sigma=r$. We assume, of course, that $r<1$, else the series will not converge. Three ways of estimating E_k suggest themselves:

(i) Apply the Corollary to Theorem 1 to the inequality

$$|E_k|_1\leqslant\sum_{m=k}^{\infty}|A^k|_1;$$

(ii) Observe that $E_k=A^k(I-A)^{-1}$, and apply Theorem 2 with $f(z)=z^k(1-z)^{-1}$.

(iii) Write $|E_k|_1\leqslant|A^k|_1|(I-A)^{-1}|_1$; estimate $|A^k|_1$ by the Corollary to Theorem 1, and the second factor using Theorem 2.

We are aiming for compactness of form as much as sharpness, so that we shall be able to judge how many terms to take using the back of an envelope. The following will afford us a simplification of our estimates.

LEMMA.

$$\sum_{i=0}^{l-1}\binom{k}{i}y^i\leqslant\frac{k^l y^l}{(l-1)!}$$

provided $ky \geqslant l + \sqrt{l}$.

Proof. Always

$$\binom{k}{i} \leqslant k^i/i!.$$

Let $a_i = k^i y^i/i!$. It will be found that $ky \geqslant l + \sqrt{l}$ is precisely the condition needed to ensure that the second differences of the sequence $(a_i)_0^l$ are nonnegative, and it then follows from graphical considerations that

$$a_1 + \cdots + a_{l-1} \leqslant l \cdot \frac{a_1 + a_l}{2} \leqslant l a_l$$

$$= \frac{k^l y^l}{(l-1)!}.$$

Method (iii): Let $g(z) = (1-z)^{-1}$. Then $g^{(j)}(z) = j!(1-z)^{-j-1}$, so that, when $|A|_\sigma = r$, $\|g^{(j)}\|_A \leqslant j!(1-r)^{-j-1}$. Hence, by Theorem 2,

$$|(I-A)^{-1}|_1 \leqslant \sum_{j=0}^{n-1} \binom{n}{j+1} |A|^j (1-r)^{-j-1}$$

$$= |A|^{-1} \sum_{j=1}^{n} \binom{n}{j} \left(\frac{|A|}{1-r}\right)^j$$

$$= |A|^{-1} \left\{ \left(1 + \frac{|A|}{1-r}\right)^n - 1 \right\}$$

$$= \frac{n}{1-r} \left(1 + \frac{x}{1-r}\right)^{n-1},$$

where $0 < x < |A|$, by the mean-value theorem, and hence

$$|(I-A)^{-1}|_1 \leqslant \frac{n}{1-r} \left(1 + \frac{|A|}{1-r}\right)^{n-1}.$$

Combining this with the Corollary to Theorem 1, we have

$$|E_k|_1 \leqslant |A^k|_1 |(I-A)^{-1}|_1$$

$$\leqslant \binom{k}{n-1} r^{k-n+1} |A|^{n-1} \frac{n}{(1-r)^n} (1-r+|A|)^{n-1}.$$

Method (ii): Let $f(z)=z^k(1-z)^{-1}$. By Leibniz's theorem

$$f^{(j)}(z)=\sum_{i=0}^{j}\binom{j}{i}k(k-1)\cdots(k-i+1)z^{k-i}\frac{(j-i)!}{(1-z)^{j-i+1}}$$

$$=j!\sum_{i=0}^{j}\binom{k}{i}\frac{z^{k-i}}{(1-z)^{j-i+1}}.$$

The power-series expansion of $f^{(j)}(z)$ clearly has all its coefficients nonnegative, so that $f^{(j)}$ attains its maximum modulus on $\{z:|z|\leqslant r\}$ at $z=r$. By the Lemma, if $k\geqslant(n+\sqrt{n})r/(1-r)$, then for $j=0,1,,\ldots,n-1$,

$$\|f^{(j)}\|_A\leqslant j!\frac{r^k}{(1-r)}j+1\sum_{i=0}^{j}\binom{k}{i}\left(\frac{r}{1-r}\right)^{-i}$$

$$\leqslant r^{k-j-1}k^{j+1}.$$

Hence, by Theorem 2,

$$|E_k|_1\leqslant\sum_{j=0}^{n-1}\binom{n}{j+1}\frac{|A|^j}{j!}r^{k-j-1}k^{j+1}.$$

The term corresponding to $j=n-1$ here is

$$\frac{|A|^{n-1}r^{k-n}k^n}{(n-1)!},\tag{12}$$

and for sufficiently large k this is the biggest term.

Method (i): This is the simplest and, as it transpires, the most effective approach. From the Corollary to Theorem 1,

$$|E_k|_1\leqslant\sum_{m=k}^{\infty}|A^k|_1\leqslant\sum_{m=k}^{\infty}\binom{m}{n-1}|A|^{n-1}r^{m-n+1}.$$

The latter expression is the $(n-1)$th derivative of a geometric series, and Leibniz's formula gives us

$$|E_k|_1\leqslant|A|^{n-1}r^{k-n}\sum_{j=0}^{n-1}\binom{k}{j}\left(\frac{r}{1-r}\right)^{n-j}.\tag{13}$$

Applying the Lemma with $y=k(1-r)/r$, we obtain:

THEOREM 5. *Let A be an $n\times n$ matrix such that $|A|_\sigma \leqslant r<1$. If $k\geqslant (n+\sqrt{n})r/(1-r)$, then the truncation error*

$$E_k=(I-A)^{-1}-(I+A+\cdots+A^{k-1})$$

satisfies

$$|E_k|_1 \leqslant \frac{|A|^{n-1}r^{k-n}k^n}{(n-1)!}. \tag{14}$$

Comparison with (12) shows that this is much better than method (ii). It also beats method (iii) on the score of neatness and the absence of a factor $(1-r)^n$ in the denominator. Method (ii) or (iii) might, however, be better if we had further information about the eigenvalues of A—say, that they lay in the left half plane.

Let us investigate what kind of information Theorem 5 yields in some concrete cases. We sum the geometric series (1) using the iterative scheme

$$B_0=I,\ B_{i+1}=B_i(I+A^{2^i}),$$

which doubles the number of terms summed at each step. Suppose that A is a 20×20 matrix of which we know that $|A|\leqslant 1$, $|A|_\sigma \leqslant \frac{1}{2}$. After p steps, $k=2^p$ and (14) tells us that

$$|E_{2^p}|_1 \leqslant \frac{2^{20p+20-2^p}}{19!}.$$

Stirling's formula gives $\log_{10}(19!)\doteqdot 17$, and so, roughly speaking,

$$\log_{10}|E_{2^p}|_1 \leqslant (20p+20-2^p)\times 0.3-17$$

This yields approximate inequalities

$$|E_{2^7}|_1 \leqslant 10^{-7}, \qquad |E_{2^8}|_1 \leqslant 10^{-39}.$$

If, for the same n and $|A|$, we know only that $|A|_\sigma \leqslant 0.9$, then we must go further:

$$|E_{2^{10}}|_1 \leqslant 10^{-7}, \qquad |E_{2^{11}}|_1 \leqslant 10^{-52}.$$

For a 50×50 matrix, $|A| \leqslant 1$, $|A|_\sigma \leqslant \frac{1}{2}$ gives, roughly

$$\log_{10}|E_{2^r}|_1 \leqslant 15p - 0.3\times 2^p - 47,$$

and so

$$|E_{2^8}|_1 \leqslant 10^{-4}, \qquad |E_{2^9}|_1 \leqslant 10^{-65},$$

while if $n=50$, $|A| \leqslant 1$, $|A|_\sigma \leqslant 0.9$, the best we can do is

$$|E_{2^{12}}|_1 \leqslant 10^{-85}.$$

It is clear that we could obtain similar results for the matrix series (2) and many other power series.

REFERENCES

1 Z. Dostál, Norms of iterates and the spectral radius of matrices, to appear.
2 Z. Dostál, Polynomials of the eigenvalues and powers of matrices, *Comment. Math. Univ. Carolinae* 19:459–469 (1978).
3 C. Foias, Contractive intertwining dilations and waves in layered media, VIth International Congress of Mathematicians, Helsinki, 1978.
4 I. C. Gohberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Translations Math. Monographs 18, Amer. Math. Soc., 1969.
5 B. Sz. Nagy and C. Foias, *Harmonic Analysis of Operators on Hilbert Space*, North Holland–Akadémiai Kiadó, Amsterdam–Budapest, 1970.
6 V. Pták, Spectral radius, norms of iterates and the critical exponent, *Linear Algebra and Appl.* 1:245–260 (1968).
7 V. Pták, A lower estimate for the spectral radius, *Proc. Amer. Math. Soc.*, to appear.
8 V. Pták, An infinite companion matrix, *Comment. Math. Univ. Carolinae* 19:447–458 (1978).
9 V. Pták and N. J. Young, Functions of operators and the spectral radius, Linear Algebra and Appl., 29:357–392 (1980).
10 R. A. Smith, Matrix equation $XA+BX=C$, *SIAM J. Appl. Math.* 16:198–201 (1968).
11 J. D. Stafney, Functions of a matrix and their norms, *Linear Algebra and Appl.* 20:87–94 (1978).
12 N. J. Young, Norms of matrix powers, *Comment Math. Univ. Carolinae* 19:415–430 (1978).
13 N. J. Young, A bound for norms of functions of matrices, *Linear Algebra and Appl.*, to appear.

14 N. J. Young, Norm and spectral radius for algebraic elements of a Banach algebra, *Math. Proc. Cambridge Philos. Soc.* 88:129–133 (1980).
15 N. J. Young, Analytic programmes in matrix algebras, *Proc. London Math. Soc.* 36:226–242 (1978).
16 N. J. Young, Norms of powers of matrices with constrained spectra, *Linear Algebra and Appl.* 23:227–244 (1979).

Received 19 December 1979; revised 25 January 1980